Language Intelligence 2024

AI Use Case Slam

Vienna, November 20

Quality Prediction of Machine Translation

Assaf Siani, Ilan Kernerman

lexicala by K DICTIONARIES

LEXICALA

- lexical resources for LSPs
- for and across 50 languages
- human-machine convergence
- mapping linguistic DNA
- multi-layer – mono-, bi-, multi-, cross-lingual – data
- cooperation with industry & academia worldwide
- MT – AI – language learning
- data-you-trust

PAPER

In **PreQuEL: Quality Estimation of Machine Translation Outputs in Advance**, Don-Yehiya et al. (2022) demonstrated that it is possible, to some extent, to predict whether a sentence has characteristics that make it easier or more difficult to translate automatically.

Shachar Don-Yehiya, Leshem Chosen, and Omri Abend. 2022. PreQuEL: Quality Estimation of Machine Translation Outputs in Advance. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11170–11183. December 7-11, 2022.*
© 2022 Association for Computational Linguistics
https://aclanthology.org/2022.emnlp-main.767.pdf

RESEARCH FEATURES

- News corpus
- Diverse Domains
- English to German
- English to Chinese
- Estonian to English
- 7K segments – training, 1K – development, 1K – test
- Translations by Fairseq
- Direct Assessment transformed to 1-0 rank

RESULTS

	en-de	et-en
PreQuEL Models		
SIMPLE	0.196±0.02	0.602±0.00
SIMPLE Aug	0.315±0.00	-
COMBINED+	0.326±0.02	-
COMBINED-	0.265±0.00	-
MULTITASK	**0.336±0.01**	-
Baseline	0.135	0.050
Upper Bound (QE)		
TransQuest	0.381±0.04	0.767±0.00

Table 1: Pearson's r of the predictions of the PreQuEL models with the DA scores, for en-de and et-en. All our models outperform the baseline, and the augmentation improves the results both as intertraining and multitask. The et-en pair correlation is much higher than the correlation of en-de.

- Positive Pearson's correlation.
- Inspired to perform such work on English to Hebrew data.

BUSINESS MOTIVATION

- Users of machine translation seek ways to reduce costs.
- If we can predict that a sentence is likely to result in poor-quality translation, we could opt not to MT it, saving both time and resources.
- This approach also creates opportunities to generate synthetic parallel corpora for low-resource languages economically.

DATA PREPARATION (1)

- We used example sentences from *Password English Learner's Dictionary* and prompted ChatGPT 4.0 to generate longer, equivalent sentences.
- The sentences were then translated with Google Translate, Bing Translator, and Yandex Translator.

DATA PREPARATION (2)

- By applying various mathematical methods to assess quality, we selected the translation that appeared to be the best among the three.
- If none of the translations met the quality standards, we excluded that sentence from the dataset.
- The resulting translations (16K segments) were manually reviewed by an expert and rated on a scale from 1 to 5.

RANKING

Translation quality assessed on a 1-5 scale

- **5.** excellent
- **4.** good; minor amendments to be made
- **3.** medium; major amendments to be made
- **2.** poor; many amendments to be made
- **1.** extremely poor

PRINCIPLES

Errors categorized according to severity

- **Minor.** inaccuracies that don't affect meaning
- **Major.** problematic inconsistencies that affect meaning or interpretation
- **Neutral.** fine, acceptable translations, need small modification due to personal preference

EVALUATION CRITERIA

- Accuracy
  - content fidelity
  - terminology consistency
- Linguistic Quality
  - grammar and syntax
  - punctuation and spelling
- Style and Tone
  - stylistic fidelity
  - cultural appropriateness

- Readability and Fluency
  - natural flow
  - clarity
- Functional Equivalence
  - purpose alignment
  - literal vs. non-literal translations

SAMPLES

4	As I walked through the museum, I marveled at the grandeur of the massive pillars supporting the ceiling.
	כשעברתי במוזיאון, התפעלתי מהפאר של העמודים המסיביים התומכים בתקרה.
3	The lifeguard quickly responded to the swimmer in distress and brought them back to shore, ensuring they were safe and sound.
	המציל הגיב במהירות לשחיין במצוקה והחזיר אותם לחוף, ווידא שהם בריאים ושלמים.
1	The small town was absorbed into the nearby city, causing its identity and local businesses to disappear.
	העיירה הקטנה נקלטה בעיר הסמוכה, מה שגרם לזהותה ולעסקים המקומיים להיעלם.
5	The government implemented strict financial regulations to ensure that banks remain solvent during times of economic instability.
	הממשלה הטמיעה תקנות פיננסיות קפדניות כדי להבטיח שהבנקים יישארו סולבנטיים בתקופות של חוסר יציבות כלכלית.

TRAINING

- We trained mBert using BertForSequenceClassification
  - multilingual (including Hebrew)
  - strong text classification capabilities
- Train – 80%, Val – 10%, Test 10%
- 15 epochs
- Learning rate – 0.00001(1e-5)

RESULTS

- Several trainings were made for both fully ranked and binary tests (we set ranking 4 and 5 to 1; otherwise 0).
- Ranked data trainings demonstrated positive Pearson's correlation **$0.25 < \rho < 0.32$**.
- Binary data training demonstrated positive Pearson's correlation **$0.35 < \rho < 0.4$**.
  - We expect that, with additional data, the results can be improved.
  - This demonstrates that we can, in fact, predict whether an English sentence is likely to be machine-translated to Hebrew with reasonable quality.

ONGOING

Research extended with University of Vienna interns to more language pairs

- English to German
- English to Polish
- English to Romanian

THANK YOU

[θæŋk juː] *interjection*

a common elliptical expression used to express gratitude or appreciation

Thank you for your attention ☺ Thank you for your cooperation!

Afrikaans **dankie**
Albanian **faleminderit**
Arabic **شُكْرا، أَشْكُرَك**
Azerbaijani **təşəkkür edirəm**
Bulgarian **благодаря**
Catalan **gràcies**
Chinese **謝謝(你)**
Croatian **hvala**
Czech **děkuji**
Danish **tak**
Dutch **dank je**
Estonian **aitäh, tänan teid**
Farsi **ممنون**
Finnish **kiitos**
French **merci**
Frisian **tanke, tankjewol**
German **danke**
Greek **(σε, σας) ευχαριστώ**
Hebrew **תּוֹדָה**
Hindi **धन्यवाद**
Hungarian **köszönöm!**
Icelandic **þakka þér**
Indonesian **terima kasih**
Italian **grazie**
Japanese **ありがとう**
Korean **감사합니다**
Latvian **paldies; pateicos**
Lithuanian **ačiū**
Malay **terima kasih**
Norwegian **tusen takk (for)**
Polish **dziękuję**
Portuguese **obrigado/-da**
Romanian **mulţumesc**
Russian **благодарю**
Serbian **hvala**
Slovak **ďakujem**
Slovenian **hvala**
Spanish **gracias**
Swedish **tack, tackar!**
Thai **การแสดงความขอบคุณ**
Turkish **teşekkür ederim**
Ukrainian **дякую; спасибі**
Urdu **آپ کا شکریہ**
Vietnamese **cảm ơn**